Cite this paper: Goode, N., Salmon, P. M., Taylor, N. Z., Lenné, M. G., & Finch, C. F. (2017). Developing a contributing factor classification scheme for Rasmussen's AcciMap: Reliability and validity evaluation. *Applied Ergonomics*, 64, 14-26.

Developing a contributing factor classification scheme for Rasmussen's AcciMap: reliability and validity evaluation

Goode, N.[1*], Salmon, P.M. [1], Taylor, N.Z. [1], Lenné, M.G. [2] & Finch, C.F. [3]

[1]Centre for Human Factors and Sociotechnical Systems, Faculty of Arts, Business and Law, University of the Sunshine Coast, Australia

[2]Monash Accident Research Centre, Monash University, Australia

[3]Australian Centre for Research into Injury in Sport and its Prevention, Federation University Australia, Australia

*Corresponding author: Natassia Goode, PhD, Email: ngoode@usc.edu.au

Cite this paper: Goode, N., Salmon, P. M., Taylor, N. Z., Lenné, M. G., & Finch, C. F. (2017). Developing a contributing factor classification scheme for Rasmussen's AcciMap: Reliability and validity evaluation. *Applied Ergonomics*, 64, 14-26.

Developing a contributing factor classification scheme for Rasmussen's AcciMap: reliability and validity evaluation

One factor potentially limiting the uptake of Rasmussen's (1997) Accimap method by practitioners is the lack of a contributing factor classification scheme to guide accident analyses. This article evaluates the intra- and inter-rater reliability and criterion-referenced validity of a classification scheme developed to support the use of Accimap by led outdoor activity (LOA) practitioners. The classification scheme has two levels: the system level describes the actors, artefacts and activity context in terms of 14 codes; the descriptor level breaks the system level codes down into 107 specific contributing factors. The study involved 11 LOA practitioners using the scheme on two separate occasions to code a pre-determined list of contributing factors identified from four incident reports. Criterion-referenced validity was assessed by comparing the codes selected by LOA practitioners to those selected by the method creators. Mean intra-rater reliability scores at the system (*M* = 83.6%) and descriptor (*M* = 74%) levels were acceptable. Mean inter-rater reliability scores were not consistently acceptable for both coding attempts at the system level (M_{T1} = 68.8%; M_{T2} = 73.9%), and were poor at the descriptor level (M_{T1} = 58.5%; M_{T2} = 64.1%). Mean criterion referenced validity scores at the system level were acceptable (M_{T1} = 73.9%; M_{T2} = 75.3%). However, they were not consistently acceptable at the descriptor level (M_{T1} = 67.6%; M_{T2} = 70.8%). Overall, the results indicate that the classification scheme does not currently satisfy reliability and validity requirements, and that further work is required. The implications for the design and development of contributing factors classification schemes are discussed.

Keywords: systems thinking; reliability; validity; incident classification

Cite this paper: Goode, N., Salmon, P. M., Taylor, N. Z., Lenné, M. G., & Finch, C. F. (2017). Developing a contributing factor classification scheme for Rasmussen's AcciMap: Reliability and validity evaluation. *Applied Ergonomics*, 64, 14-26.

Developing a contributing factor classification scheme for Rasmussen's AcciMap: reliability and validity evaluation

1. Introduction

As evidenced by a recent special issue of *Applied Ergonomics* that included 18 papers dedicated to the work of Jens Rasmussen (Waterson, Jenkins, Salmon, & Underwood, 2016), Rasmussen's (1997) risk management framework and Accimap methodology are becoming increasingly prominent in safety research. However, whilst there are now many peer reviewed articles that have applied both to accidents from a wide variety of domains (Waterson et al., 2016), the method does not yet appear to have achieved the same popularity with practitioners as has the Human Factors Accident Classification System (HFACS; Wiegmann & Shappel, 2003). HFACS has been used across a wide variety of industries (e.g. aviation, defence, maritime, rail) to underpin investigations, analyse existing data and underpin incident reporting systems, and a number of organisations have developed domain-specific variants (Baysari, McIntosh, & Wilson, 2008; Chauvin, Lardjane, Morel, Clostermann, & Langard, 2013; O'Connor, 2008; Olsen, 2011; Olsen & Shorrock, 2010; Shappell & Wiegmann, 2012; Walker, O'Connor, Phillips, Hahn, & Dalitsch, 2011; Wiegmann & Shappell, 2001). In contrast, Accimap has been applied in relatively few organisations, and applications appear to be limited to single major incident analyses (e.g. the Australian Transport Safety Board; Underwood & Waterson, 2013a).

One of the factors potentially limiting the uptake of Accimap by practitioners is that it does not provide a classification scheme of contributing factors to guide the analysis (Salmon, Cornelissen, & Trotter, 2012; Underwood & Waterson, 2013b). This creates concerns regarding its reliability and validity. In contrast to application of HFACS and its domain-specific variants, analysts must use a "bottom-up" coding approach to identify contributing factors and relationships from the data. While this makes the technique easy to apply to new domains, it also means that the analysis is dependent on the subjective judgment of the analyst (Salmon et al., 2012; Underwood & Waterson, 2013b, 2014). This makes it difficult to produce useful summaries of multiple incidents and prevents Accimap's implementation within incident reporting systems. This article reports on the evaluation of a contributing factor classification scheme that was developed to support the use of Accimap as part of an incident reporting system called UPLOADS (Understanding and Preventing Led Outdoor Accidents Data System, see Salmon et al., 2016). This work builds on earlier research that developed a standardised format and procedure for Accimap analyses of single accidents (Branford, Hopkins, & Naikar, 2009). The purpose of the contributing factor classification scheme reported on in this article is to standardise the aggregation of Accimap analyses across multiple incidents.

There are various characteristics that need to be met in order for a classification scheme to prove useful in accident analysis and prevention efforts (Wallace & Ross, 2006b). As with all ergonomics methods, two important features are, first, that the same analyses are produced by the creators of the method and (trained) intended end users (i.e. criterion-referenced validity), and second, that the method is reliable when used on different occasions (by the same analysts) and by different analysts (i.e. intra-rater and inter-rater reliability; Stanton, 2016; Stanton & Young, 1999, 2003). Intra-rater reliability is important for contributing factor classification schemes because incident data is typically collected and analysed over months and years – so multiple analyses of the data will be undertaken during this time (Olsen & Shorrock, 2010). Inter-rater reliability is important where multiple analysts or safety departments contribute to data coding, and when this is high it indicates that a classification scheme is logically organized and parsimonious (Ross, Wallace, & Davies, 2004). Evaluating criterion-referenced validity and intra- and inter-rater reliability allows for an initial assessment of whether a method actually works in practice, and gives practitioners a basis for

judging the potential utility of applying the method within their organisation (Stanton, 2016; Stanton & Young, 1999, 2003). Despite the importance of evaluating ergonomics methods in this fashion (e.g. Baysari, Caponecchi, & McIntosha, 2011; Cornelissen, McClure, Salmon, & Stanton, 2014; Kirwan, 1997; Stanton & Young, 1998; Stanton et al., 2009; Stanton & Young, 2003) it has rarely been undertaken (Stanton, 2016).

The aim of the present study was to evaluate the reliability and validity of a contributing factor classification scheme that was designed specifically for the Accimap method. The study was conducted with the intended end users of the scheme, safety practitioners from the led outdoor activity (LOA) domain. A test-retest paradigm was used to assess intra- and inter-rater reliability. Validity was assessed by comparing how practitioners and the method creators assigned codes to classify contributing factors. The following sections first present a brief overview of the context for this research, the LOA domain, and the methodological issues that informed the design of this study. This is followed by a description of Rassmussen's (1997) risk management framework and Accimap technique, and the initial development of the contributing factor classification scheme before the evaluation results are presented.

1.1. The research context: the led outdoor activity domain

"Led" outdoor activities are formally defined as facilitated or instructed activities within outdoor education and recreation settings (Salmon, Williamson, Lenne, Mitsopoulos-Rubens, & Rudin-Brown, 2010). Examples include activities such as bushwalking, canyoning, kayaking, rock climbing and camping.

Over the past decade, the domain has experienced high profile fatalities that have highlighted the systemic nature of LOA incidents (Salmon et al., 2012; Salmon et al., 2010). For example, six students and their teacher died while on a gorge walking activity in New Zealand in 2008. The coroner and an independent investigation highlighted multiple contributing factors relating to the instructor, her manager, the activity centre, the local weather service, the auditing system, and government legislation and regulation (Brookes, Smith, & Corkill, 2009; Davenport, 2010). Subsequent research examining other fatal LOA accidents (Salmon et al., 2010) and more common, but less catastrophic, injuries (e.g. Salmon, Goode, Lenné, Finch, & Cassell, 2014) revealed similar findings, showing how multiple factors across the "led outdoor activity system" contribute to incidents. Prior to these analyses, research examining incident causation in this domain had focused on the immediate context of the incident (e.g. activity leader knowledge of environmental hazards and experience, supervision, weather) (e.g. Brookes, 2003, 2004; Curtis, 1995; Haddock, 2004; Hogan, 2002), with little acknowledgement of the factors at the higher levels of the system. The need for a LOA-specific systems thinking-based approach to incident analysis and prevention was therefore identified (e.g. Salmon et al., 2012; Salmon et al., 2010).

The contributing factor classification scheme described in this article was therefore developed to support the application of Accimap as part of an incident reporting system to be used by LOA organisations. The scheme is intended to be used by practitioners to guide their incident investigations and is also implemented within an incident reporting system, known as UPLOADS (Understanding and Preventing Led Outdoor Accidents Data System, see Salmon et al., 2016). The intended end users of the scheme are practitioners from a diverse range of organisations including not-for-profit outdoor education and recreation providers, outdoor education departments within schools, school camps, adventure tourism operators and outdoor therapy programs. These organisations range in size from large organisations with many hundreds of staff to sole operators, and are distributed across Australia, often in remote areas (Service Skills Australia, 2010, 2013;

Williams & Allen, 2012). These characteristics were a key consideration in developing the contributing factor classification scheme and the associated training material (Goode, Finch, Cassell, Lenne, & Salmon, 2014a).

1.2. Reliability and validity studies

There are a number of key methodological issues that need to be considered when designing reliability and validity studies.

The first consideration is the presentation of the study materials. In some studies, participants use a checklist to select all the codes that they feel apply to an incident, without specifying the contributing factors they have identified (e.g. O'Connor, 2008). This design is not sufficient for testing whether codes are assigned consistently applied to the same types of factors, as participants may select the same codes but identify quite different contributing factors (Ross et al., 2004). In other studies, participants are required to identify the factors that contribute to the incident and select codes to describe these factors (e.g. Gordon, Flin, & Mearns, 2005). This design is also problematic, as low inter-rater reliability may be attributable to disagreements over the relevancy of the factor, the selection of the code, or both (Olsen, 2013; Ross et al., 2004; Wallace & Ross, 2006a). An alternative approach is to present incident reports with a list of pre-identified contributing factors along with a space beside each for coding (e.g. Olsen, 2011). This allows for an examination of whether participants consistently use the same codes to describe the same contributing factors. This design was used in the current study as the aim of study was to evaluate the reliability and validity of the selection of codes by practitioners in the LOA domain.

Second, the type and extent of training given to participants should be explicitly justified with reference to the intended context of use. While some have argued that high inter-rater reliability for coding frameworks is only reached by giving extensive face-to-face training (Krippendorff, 2004), a recent review of incident classification scheme reliability studies did not find any relationship between inter-rater reliability and providing training by email, workbook or face-to-face group sessions (Olsen, 2013). However, any study design where participants are allowed to discuss the selection of codes, ask the method creators questions, or receive feedback on coding will potentially result in improvements in inter-reliability as it allows for calibration between participants (Olsen, 2013). Conversely, this type of study design is likely to have a much smaller impact on intra-rater reliability, and may even have a negative impact if feedback or clarification is provided between repeated codings. We argue that the training format should be selected based on the resources available within the intended context of use. For example, incident classification schemes designed to be used by safety teams in large organizations might reasonably involve more training than those designed to be used in small businesses without dedicated safety personnel. During the initial design phase of the classification scheme, minimising workload was rated as the highest priority by LOA practitioners (Goode et al., 2014a). Face-to-face training was considered not practical, as the end user organisations are distributed across Australia. Therefore, the study was conducted via email and written training material was provided.

A third, related issue, is the type of participants used in the evaluation. Wallace and Ross (2006a) recommend involving those who will use the scheme in the "real world" (p.243). While this makes sense intuitively, access to the intended end users may be limited; therefore researchers often make use of human factors specialists, other researchers and students (Olsen, 2013). This, of course, leads to criticisms of poor ecological validity. In relation to ergonomics methods more generally, Stanton (2016) argues that a method is "considered to have minimally acceptable reliability if the method's expert creator could achieve repeatable results on different occasions. At the other extreme would be a method that delivered the same results when used by anyone with even a little training" (p. 347). As the development of classification schemes involves an iterative process of testing and

refinement, this suggests that the early stages can involve the research team or human factors specialists, while later stages of development should involve the intended end users. The classification scheme in this study is in its later stages of development, therefore a sample of intended end users was involved.

Fourth, the method used to calculate reliability and validity remains a point of much contention in the literature. The most commonly used methods are the percentage of participants who agree on a code (e.g. O'Connor, 2008), the Index of Concordance (also referred to as raw agreement; e.g. Olsen & Shorrock, 2010), Kappa (e.g. Makeham et al., 2008) and the signal detection paradigm introduced by Stanton and colleagues (Baber & Stanton, 1994; Cornelissen et al., 2014; Stanton et al., 2009; Stanton & Stevenage, 1998; Stanton & Young, 1999, 2003). The percentage of participants who agree on a code has been widely criticised primarily as it does not take into account differences in the total number of codes selected by each participant (e.g. Olsen & Shorrock, 2010; Ross et al., 2004; Wallace & Ross, 2006b). In comparison, the Index of Concordance accounts for the number of agreements and disagreements for each pair of codes assigned by each pair of participants (Ross et al., 2004). This approach penalises classification schemes with a large number of overlapping categories, which is consistent with the definition of poor reliability (Fleishman, Quaintance, & Broedling, 1984). Kappa attempts to correct the Index of Concordance for “chance agreement” (Cohen, 1960). However, a number of authors have argued that the process of coding incident reports typically violates the basic assumptions of Kappa (i.e. the items to be coded are independent; the categories are independent, mutually exclusive and exhaustive; and the coders operate independently; Cohen, 1960).

An alternative approach is the signal detection paradigm, which allows for an assessment of the sensitivity of the method (Stanton, 2016; Stanton & Stevenage, 1998). In studies assessing criterion-referenced validity, this approach can be used to assess the concordance between expert and novice coding, by calculating the number of “hits”, “misses”, “false alarms” and “correct rejections” across the contributing factor classification scheme. Hits represent the number codes that were selected by both the expert and novice. Misses represent the codes that were selected by the expert but not by the novice. False alarms represent the codes that were selected by the novice but not by the expert. Correct rejections represent the codes from the classification scheme that were not selected by either novice or expert. An index of sensitivity can then be calculated combining these metrics (see Stanton & Stevenage, 1998). One drawback of this approach is that sensitivity can be artificially inflated by a large number of correct rejections. This is particularly problematic in studies evaluating classification schemes with large numbers of codes where only a few codes apply to each incident (e.g. O'Connor, 2008). In these cases, it is much easier to reject a large number of codes that clearly do not apply to an incident, than to correctly identify the codes that do apply (O'Connor, 2008). In designing this study, a large number of correct rejections was anticipated as the contributing factor classification scheme comprised 107 codes. This also means that the likelihood of “chance agreement” could be considered relatively low. Therefore, the Index of Concordance was used to calculate reliability and validity.

A final consideration is identifying a criterion for acceptable levels of reliability and validity, as there is no universally accepted measure. A recent review found that, across 25 studies, the average value used to indicate acceptable percentage agreement was 76%, with a range of 70% to 88% (Olsen, 2013). One problem with using the average from previous studies is that percentage agreement is highly influenced by the sample size. It is much easier to obtain a score above 76% in a study involving four participants (e.g. Olsen, 2011), as opposed to a study involving over 100 participants (e.g. O'Connor, 2008). An alternative approach is to justify the criterion based on the context of

intended end use, with reference to the error that would be acceptable to practitioners. In this study, a criterion of 70% agreement was adopted as a reasonable minimum, in accordance with Wallace and Ross (2006b) and evaluations of similar contributing factor classification schemes (e.g. Olsen & Shorrock, 2010).

1.3. Rasmussen's (1997) risk management framework and Accimap

In line with other systems based accident analysis methods (e.g. STAMP, Leveson, 2004; HFACS; Wiegmann & Shappell, 2001), Rasmussen's risk management framework represents socio-technical systems as hierarchies comprising multiple levels, as shown in shown in Figure 1. Accimap is then used to graphically represent how the conditions, and decisions and actions of various actors across the levels of the system interact with one another to create the incident under analysis. One of the advantages of this method is that it provides a detailed representation of incident trajectories, incorporating contributory factors and their interrelations across the overall system. As mentioned, a limitation of the method is that aside from the hierarchical levels the analyst is given no guidance on: what kinds of contributory factors should be considered; at which level they should be placed; and which contributory factors are related with one another. The classification scheme that is the subject of this paper provides a set of codes for categorising the contributing factors by type, actor, and level, and the relationships between them, as an attempt to overcome some of these limitations.

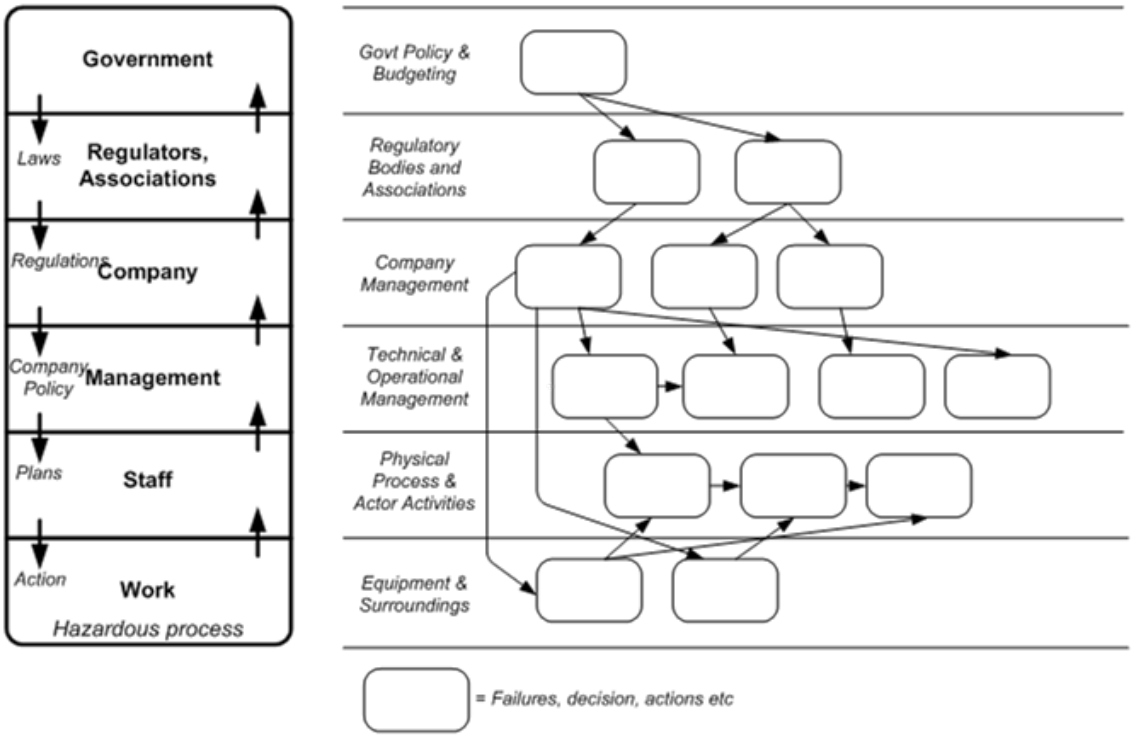

Figure 1 Rasmussen's Risk Management Framework and Accimap method (adapted from Rasmussen, 1997).

1.4. Development of the contributing factor classification scheme

Development of the classification scheme involved adapting Rasmussen's framework to describe the "LOA system" and the identification of a set of prototype codes to populate the levels. The prototype codes were based on: 1) a literature review of the contributing factors that have been identified as playing a role in LOA incidents (Salmon, Williamson, Lenne, Mitsopoulos-Rubens, & Rudin-Brown, 2009); 2) case study analyses of fatal LOA incidents (Salmon et al., 2012; Salmon et al.,

2010); and 3) analyses of existing LOA incident data (Salmon et al., 2014). The intention was to develop a classification scheme with enough detail so that aggregate incident could immediately be used to understand the specific factors that are recurring across incidents without further coding, and also show reasonable levels of reliability and validity when used by practitioners with minimal training. These two requirements needed to be balanced against each other, as highly detailed categories typically show poor levels of reliability and validity (Finch et al., 2012; O'Connor, 2008).

The prototype classification scheme is shown in Figure 2. The prototype used a hierarchical structure that is a typical feature of many incident classification schemes, such as HFACS and its variants (O'Connor, 2008; Olsen & Shorrock, 2010; Shappell & Wiegmann, 2012). The prototype had three levels of codes: the system level, the descriptor level and the specific code level. The system level described the 'led outdoor activity system' in terms of 10 codes describing the activity context; the key people involved in the activity; and the people and agencies that impact on how the activity is run. The descriptor level broke the first level categories down into between 2 and 13 descriptive categories, with a total of 55 codes. The highly specific level broke the second level categories down into between 2 and 19 highly specific factors, with a total of 325 codes. For example, Figure 3 shows how the descriptive category "Activity Equipment" was broken down into seven highly specific codes. The highly specific codes reflected the level of detail that was identified in pre-existing incident reports (Salmon et al., 2014). However, it was anticipated that reliability for this level of detail would likely be poor based on findings from evaluations of other classification schemes (e.g. O'Connor, 2008; Olsen & Shorrock, 2010). Therefore, an initial evaluation study was conducted to determine the appropriate level of detail for the classification scheme.

Cite this paper: Goode, N., Salmon, P. M., Taylor, N. Z., Lenné, M. G., & Finch, C. F. (2017). Developing a contributing factor classification scheme for Rasmussen's AcciMap: Reliability and validity evaluation. *Applied Ergonomics*, 64, 14-26.

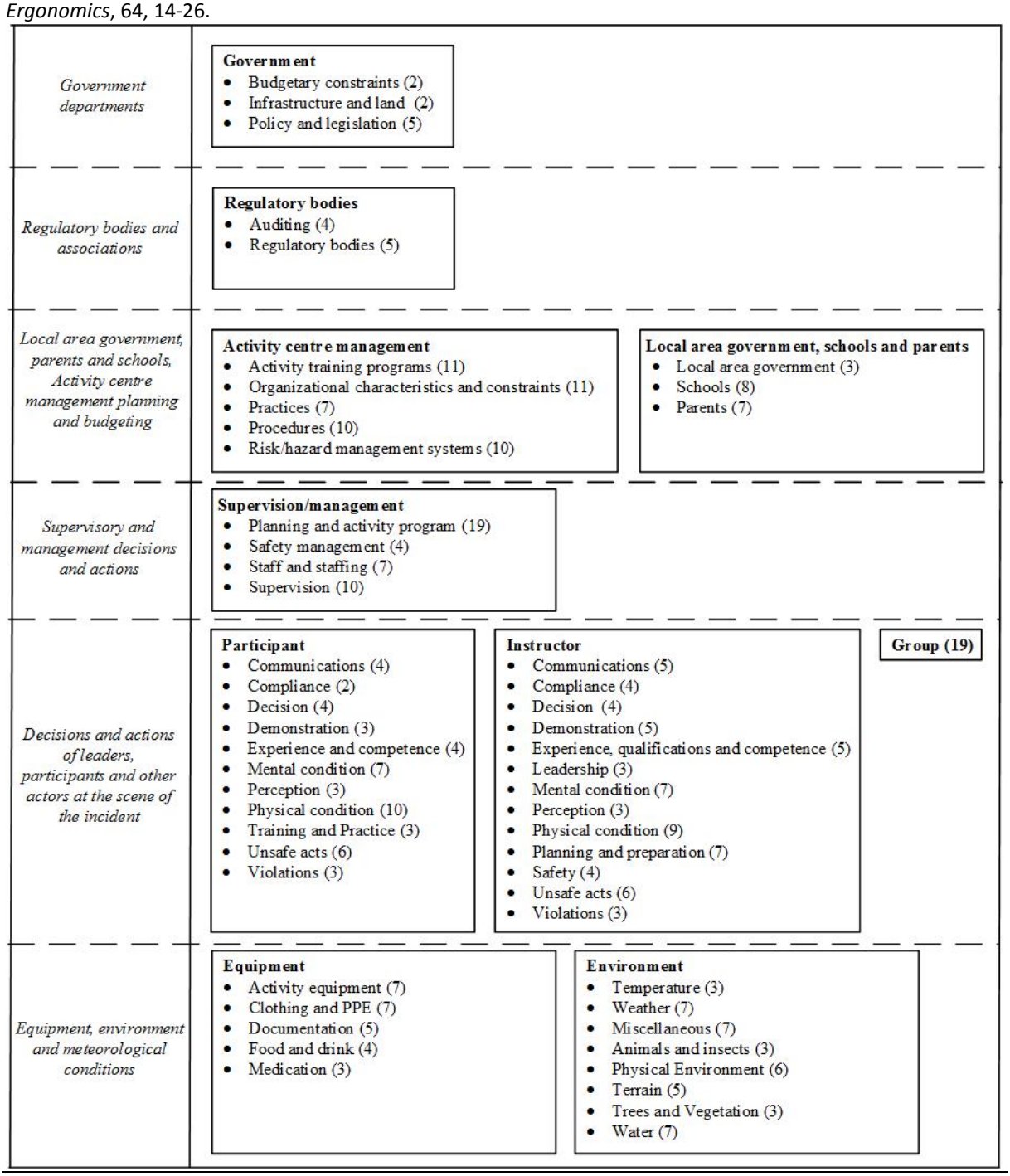

Figure 2 Overview of prototype classification scheme in the context of the adapted Accimap framework. System level codes are shown in bold. Descriptor level codes are presented below each corresponding system level factor. The number of corresponding highly specific codes are indicated in brackets.

Cite this paper: Goode, N., Salmon, P. M., Taylor, N. Z., Lenné, M. G., & Finch, C. F. (2017). Developing a contributing factor classification scheme for Rasmussen's AcciMap: Reliability and validity evaluation. *Applied Ergonomics*, 64, 14-26.

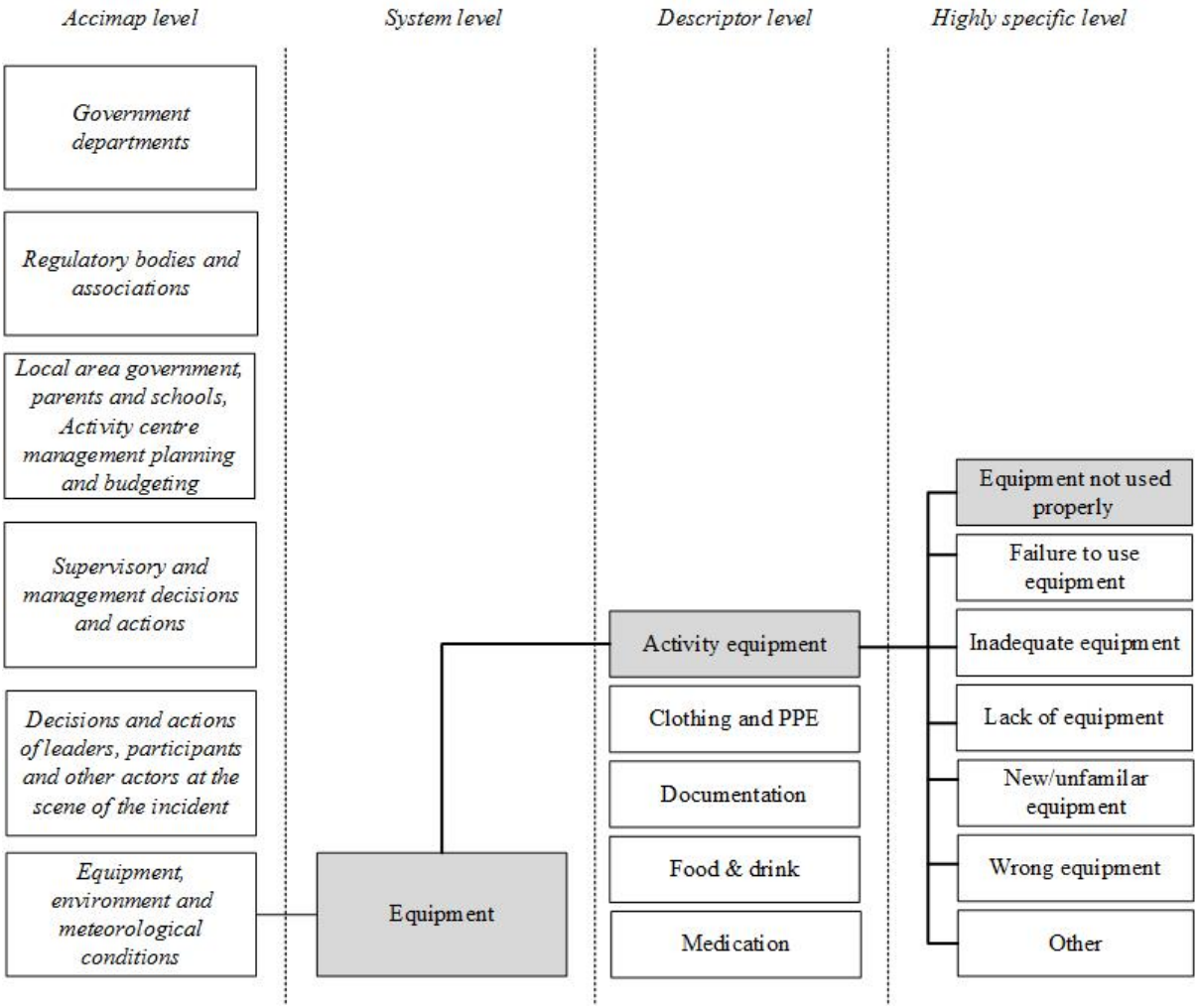

Figure 3 Example of the heirachical structure of the prototype taxonomy, with an example of the level of detail captured by the highly specific factors within the taxonomy.

This early evaluation found that the inter-rater reliability of the prototype classification scheme was very low when used by LOA practitioners to code detailed incident reports (Goode, Salmon, Lenne, & Finch, 2014b). The findings showed that only the system level showed acceptable levels of inter-rater reliability, and that participants often selected many highly specific codes to describe the same factor. In addition, many participants reported that it required over an hour to code each incident due to the large number of highly specific codes. It was concluded that the descriptor level codes might represent a reasonable trade-off between detail and reliability, although reliability at this level of detail was still poor and revisions was required to ensure the codes were discrete and parsimonious. The incident classification system was subsequently revised in light of these findings.

The revised classification scheme is presented in Figure 4, and consists of two levels of codes: the system level and the descriptor level. The system level describes the 'led outdoor activity system' in terms of 14 codes describing the actors, artefacts and activity context. The descriptor level describes specific contributing factors relating to each of these components, with a total of 107 codes.

Cite this paper: Goode, N., Salmon, P. M., Taylor, N. Z., Lenné, M. G., & Finch, C. F. (2017). Developing a contributing factor classification scheme for Rasmussen's AcciMap: Reliability and validity evaluation. *Applied Ergonomics*, 64, 14-26.

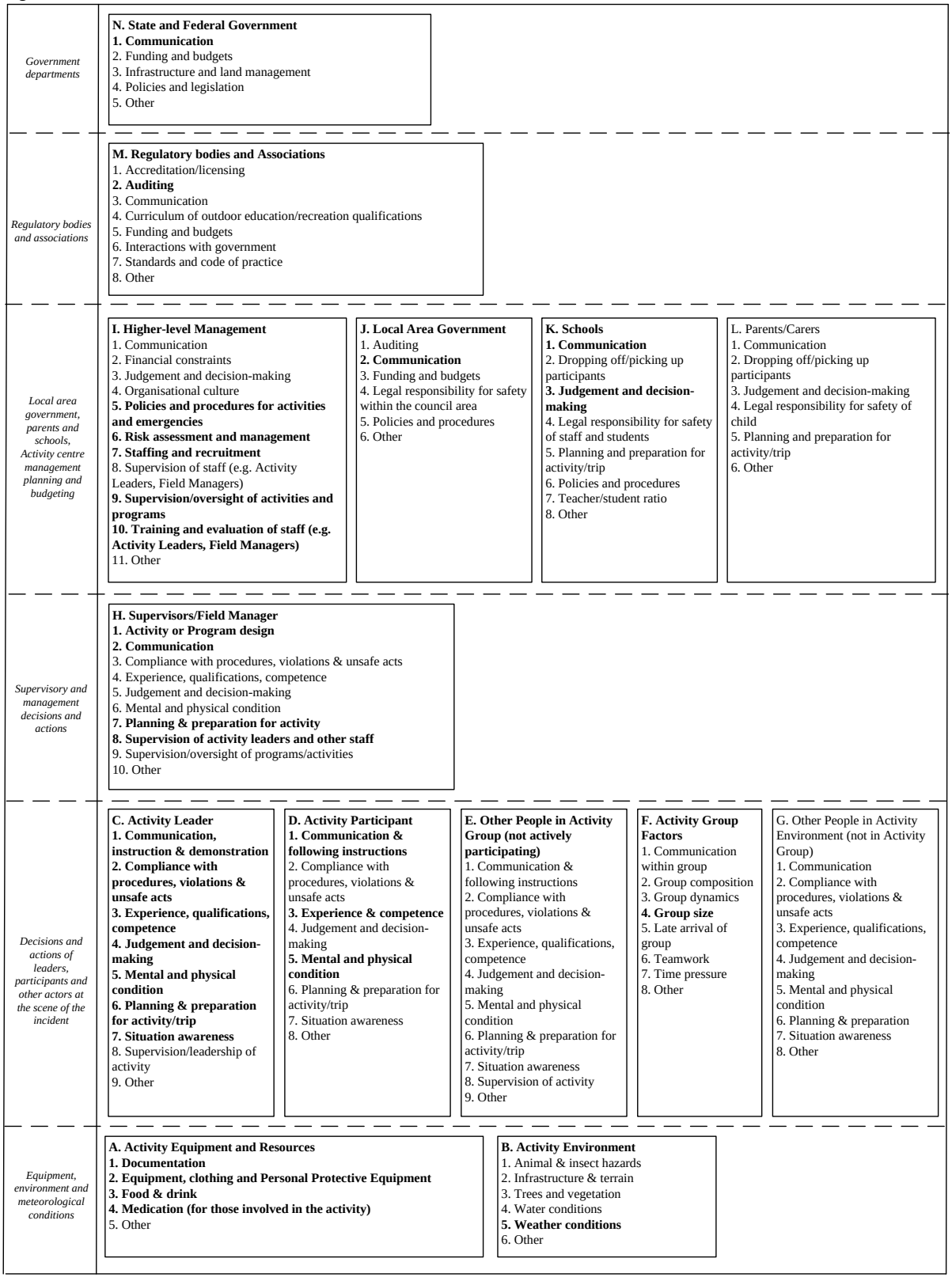

Figure 4 Overview of revised classification scheme. System level codes are numbered with letters. Descriptor level codes are numbered below each corresponding system level code. The codes highlighted in bold were used by the method creators to code the reports that are the focus of the present study.

An initial pilot of the revised classification scheme was conducted with participants with extensive experience with human factors methods (Taylor, Goode, Salmon, Lenne, & Finch, 2015). While the average inter-rater reliability at the system and descriptor levels was greatly improved from the prototype, a number of codes at both levels appeared to be hard to distinguish between and were associated with poor inter-rater reliability. To address these identified limitations, a number of clarifications were inserted into the training material to assist in distinctions between codes. Specifically, further explanation was provided for each category at the system level, and an extensive set of examples including key words were developed for each. Further evaluation of the revised classification scheme and training material is now required to establish whether they show reasonable levels of validity and reliability when used by LOA practitioners who have little experience in human factors methods or coding qualitative data. This is the focus of the remainder of this paper.

1.5. Objective of the study

In summary, the aim of this study is to evaluate the intra- and inter-rater reliability and validity of a contributing factor classification scheme when used by the intended end users, LOA practitioners. Reliability and validity will be assessed for coding at the system level and descriptor level of classification scheme.

2. Method

2.1. Design

A test-retest study design was used. This involved LOA practitioners using the classification scheme on two separate occasions (Time 1 and Time 2) to code a pre-determined list of contributing factors identified from the same pre-selected, indicative four incident reports. The period between analyses ranged from 1 to 3 months depending on when participants returned the study materials. Ethics approval was granted by the University of the Sunshine Coast Human Ethics Committee (A/14/604).

2.2. Recruitment

Managers and staff who, at the time of recruitment, played a key role in risk management within LOA organizations were invited to participate through email lists maintained by the research team and LOA professional associations (see acknowledgements). The aim was to recruit at least six to eight participants, as per Wallace and Ross (2006a) recommendations.

Seventeen people volunteered to participate and were emailed the study materials. Twelve people completed the first component of the study, and eleven of these completed the second, representing a response rate of 71% and 65% respectively. Only the data for participants who completed both components of the study were included in the analysis.

2.3. Sample

The eleven participants were primarily male (n = 10), with an average age of 42 years (SD = 5.74, range 31 to 48). On average, participants had 17 years' experience in the LOA sector (SD = 6.48, range 5 to 25 years). All participants held a managerial role in their organisation with direct responsibility for safety management (e.g. risk manager, program manager, senior teacher, director of outdoor education).

2.4. Training material

The training manual presented a brief overview of the underpinning theoretical framework (e.g. Rasmussen, 1997), and described each level of the classification system. The decision-gate style of the system was described. First, system level codes were then presented, with examples of the actors that would be categorised with each code. For example, "Other People in Activity Group"

describes the people in the activity group who are not participating in the activity but who contribute to the immediate supervision and provision of the activity e.g. parents, teachers, drivers, cooks, cleaners." Descriptor level codes were presented with extensive examples of factors that would be assigned to each code. For example, "Supervisors/Field manager: Activity or Program design" was presented with the examples: poorly designed activity; too many activities were scheduled during program; activities were inappropriate for participant skill level. These examples were taken from the analysis of over 1000 incident reports which was used by the method creators to develop the classification scheme (Salmon et al., 2014). These examples were therefore reflective of the method creators' knowledge of the scheme.

Finally, a number of clarifications were presented in the manual based on issues identified in previous evaluations (Goode et al., 2014b; Taylor et al., 2015). First, participants were instructed to only use the information provided in the reports to code the factors, and not personal experience. Second, they were directed that all issues with equipment should be coded to "Equipment, clothing and personal protective equipment." If there was explicit evidence that this failure could be attributed to the actions of a particular actor (e.g. Activity Leader or Activity Participant), then they should choose an appropriate second code. Third, participants were instructed that, although the method creators had tried to reduce the overlap between the codes as much as possible, if participants perceived that more than one code could be used to describe a particular factor, then they should select all of these codes. This final instruction was included to determine whether further distinctions between codes are required.

2.5. Content of coding booklet

The first section of the coding booklet asked participants for basic demographic information (e.g. age, gender, organisational role, and experience in the LOA sector).

The second section of the booklet contained four detailed incident reports, approximately half to a full page in length (single spaced). Only four incidents were presented in this study, as many participants in an earlier reliability study involving 10 reports had complained about the length of time required to complete the tasks (Goode, Salmon, Lenne, & Finch, 2014c).

The reports were developed to describe a range of contributory factors from the incident classification system. The reports were adapted from the Australian Accident Register, an online publicly accessible voluntary report database of accidents and serious near misses from the LOA community (https://groups.google.com/forum/#!forum/australian-accident-register). Any identifying information that referred to organisations, people or locations was changed to reflect their roles to protect the privacy of the people involved. Some details in the reports were extended to provide further details of potential contributing factors for the purpose of the coding exercise. The four reports described: (1) An outdoor rock climbing activity with adult participants which resulted in a serious injury to an Activity Leader; (2) A canyoning activity involving a large inexperienced adult group which resulted in an injury to a participant, an overdue group and a mass rescue operation; (3) A student on a school camp with a severe asthma attack, which required an emergency evacuation; and (4) An orienteering activity with a school group involving the drowning of a student.

After each report, a list of contributing factors identified by the first and second authors as contributing to the incident was presented, with a space to list the codes from the scheme that best described the factor. Table 1 shows the number of contributing factors presented after each incident, and the codes used by the method creators to classify them. Based on the codes used by the method creators to classify the contributing factors, the four reports covered 11 out 14 system

level codes, and 30 out 107 descriptor level codes; this is represented on the overview of the classification scheme presented in Figure 4.

Table 1 Summary of the number of contributing factors presented after each incident and the codes used by the method creators to classify them. The description of each code is presented in Figure 4.

Incident	No. of contributing factors presented	Codes used by method creators to classify the contributing factors
1	13	A2, A2, C2, C5, B5, C5, C4, D1, D3, H1, H8, I7, H8, I7
2	14	A2, A2, D3, C1, C3, F4, H2, A1, I5, C4, D5, D1, C6, I5, N1
3	8	D5, A4, D1, C4, K3, I9, I5, K1
4	15	A2, C1, C3, B3, J2, I6, C7, I5, A2, I5, C3, I10, H7, A1, I5, M2

2.6. Procedure

Only the third author, who was not involved in the development of the method, had contact with participants during the study. Participants completed an online form to indicate their interest in participating in the study. All subsequent correspondence was conducted via email.

For the first coding (Time 1), the training manual and coding booklets were emailed to participants, with a request to email the completed booklets back to the same email address within a fortnight. Participants received three email reminders at weekly intervals. If they did not respond, they were not included in the second phase of the study. Consent was indicated by returning the coding booklets.

For the second coding (Time 2), the same training material and coding booklet was emailed a month after they had returned their first coding booklet, with a request to return the completed analyses within a fortnight. They were instructed not to refer to their earlier analyses and to read through the training manual again. Participants again received three email reminders to return the coding booklets.

Participants did not receive any feedback on the accuracy of their coding at any stage during the study. No clarifications regarding the training manual or classification scheme were provided.

2.7. Method creator coding

The method creators (the first and second author) independently completed the coding booklets. There was approximately 95% agreement between them regarding the selection of codes across the four incidents. There was disagreement over two factors where the second author perceived that two codes applied to each factor. It was decided to include these codes in the criterion for the validity assessment.

2.8. Data analysis

The Index of Concordance (see Wallace & Ross, 2006b) was used to calculate percentage agreement at the system level and descriptor level of the classification scheme. For each pair of coders, this involved scoring agreement or disagreement for each pair of codes assigned to each contributing factor in the pre-determined list for each incident. For each pair of coders, the total number of agreements was then divided by the total number of agreements and disagreements (agreements/(agreements + disagreements). The mean was then calculated across all pairs of coders.

Agreement was assessed at both the system level and descriptor level for each pair of codes. For example, if Participant 1 selected the code “ACTIVITY EQUIPMENT AND RESOURCES: Equipment,

clothing and Personal Protective Equipment" and Participant 2 selected the code "ACTIVITY EQUIPMENT AND RESOURCES: Other" to describe the contributing factor "Activity Leader forgot to wear helmet", then they would be scored "Agreement" at the system level, and "Disagreement" at the descriptor level. If Participant 2 chose an additional code to describe the contributing factor, then they would receive a "Disagreement" for both levels for this code. For the system level and the descriptor level, the proportion of agreeing pairs of codes out of all possible pairs of codes was then calculated for each pair of coders and converted to a percentage (i.e. number of agreements divided by the number of agreements and disagreements, multiplied by 100).

To assess intra-rater reliability, the codes selected by each participant at time 1 and time 2 were compared. The Index of Concordance was then calculated for each incident.

To assess inter-rater reliability, the codes selected by each participant were compared to all other participants for time 1, and again for time 2. The mean Index of Concordance across all pairs of participants was then calculated for the four incident reports for time 1 and time 2.

To assess validity, the codes selected by each participant were compared with those selected by the method creators for time 1, and again for time 2. The mean Index of Concordance across all participants was then calculated for the four incident reports for time 1 and time 2.

A criterion of 70% was used to evaluate whether the validity and reliability of the classification scheme was acceptable.

3. Results

3.1. Number of codes selected for each incident

The number of codes selected by the method creators and participants to describe the contributing factors involved in each incident is shown in Table 2. Overall, the number of codes selected is similar to the number of contributing factors presented for each incident. This provides an initial indication that the majority of codes are distinct from one another. On average, participants selected a similar number of codes as the method creators, however, participants tended to select more codes on the first coding attempt. This indicates that the application of the codes may become clearer with practice.

Cite this paper: Goode, N., Salmon, P. M., Taylor, N. Z., Lenné, M. G., & Finch, C. F. (2017). Developing a contributing factor classification scheme for Rasmussen's AcciMap: Reliability and validity evaluation. *Applied Ergonomics*, 64, 14-26.

Table 2 Number of codes selected by the method creators and participants for the four incidents at time 1 and time 2

Coder	Incident 1 (13 factors)		Incident 2 (14 factors)		Incident 3 (8 factors)		Incident 4 (15 factors)	
	Time		Time		Time		Time	
	1	2	1	2	1	2	1	2
Method creators	14	14	14	14	8	8	16	16
1	15	15	15	15	8	8	15	-
2	13	13	14	14	8	8	15	16
3	13	13	16	14	8	8	20	17
4	16	13	14	14	9	10	15	15
5	15	13	21	15	8	8	22	16
6	13	13	14	14	8	9	15	17
7	16	18	19	17	9	9	19	19
8	13	13	14	14	8	8	15	15
9	17	16	18	16	10	9	19	21
10	13	13	15	14	8	8	18	16
11	15	13	17	14	8	8	17	15

3.2. Intra-rater reliability

A summary of the intra-rater reliability scores for each participant, along with the days between the first and second coding attempts, is shown in Table 3. The results show that the mean Index of Concordance between the two coding attempts was above the acceptable threshold for the system level and descriptor level codes. There was no obvious relationship between the days between coding attempts and the mean Index of Concordance. At the individual level, all participants apart from Participant 4 scored above the acceptable threshold at the system level. Seven out of 11 participants scored above the acceptable threshold at the descriptor level. Participant 4's scores in comparison to the sample indicate they are potentially an outlier; therefore they were excluded from further analyses.

Table 3 Summary of intra-rater reliability scores: mean Index of Concordance (%) for system level and descriptor level codes assigned by the same participants across the four incidents at time 1 and time 2

Participant	Days between Time 1 and Time 2	System level (%)	Descriptor level (%)
1	34	93.4	80.2
2	62	89.4	80.6
3	43	94.6	92.7
4	58	69.3	56.9
5	38	79.0	67.9
6	79	74.4	72.2
7	78	79.7	61.7
8	77	88.4	90.9
9	48	81.8	73.4
10	28	83.2	71.9
11	51	86.3	65.4
Mean overall	54	83.6	74.0

3.3. Inter-rater reliability

Descriptive statistics for the inter-rater reliability scores for the system level codes and descriptor level codes are shown in Table 4. At the system level, the results show that the mean Index of Concordance was slightly below the acceptable threshold for time 1, and slightly above for time 2. At the descriptor level, the results show the mean Index of Concordance was substantially below the acceptable threshold for both coding attempts. However, at both levels, the large standard deviations indicate that was also a wide variation in agreement amongst participants.

Overall, there were many system and descriptor level codes where there was a lack of agreement between participants. Two consistent difficulties were identified: identification of the appropriate actors at the higher levels of the framework; and attribution of codes to factors relating to activity equipment and resources. In relation to the first issue, participants often disagreed about whether a factor related to "supervisors/field manager", "higher-level management" or "schools". In relation to the second issue, some participants assigned "lack of equipment/medication/documentation" to the actor responsible (e.g. activity participant or leader), rather than to the object.

Table 4 Descriptive statistics for inter-rater reliability scores: mean (M) Index of Concordance (%) between coders for system level and descriptor level codes for each incident at time 1 and time 2

	System level				**Descriptor level**			
	Time				Time			
	1		2		1		2	
Incident	M	SD	M	SD	M	SD	M	SD
1	73.1	9.2	75.3	11.1	57.2	11.5	64.2	11.7
2	66.0	11.9	78.2	9.5	58.2	12.8	64.1	10.2
3	73.8	11.6	71.6	15.3	56.8	15.2	61.9	13.5
4	62.2	9.1	70.3	8.8	61.7	10.6	66.4	10.5
Mean overall	68.8	10.5	73.9	11.2	58.5	12.5	64.1	11.5

3.4. Validity

Descriptive statistics for the validity scores for the system level codes and descriptor level codes are shown in Table 5. At the system level, the results show that the mean Index of Concordance was above the acceptable threshold for time 1 and time 2. At the descriptor level, the results show that the mean Index of Concordance was slightly below the acceptable threshold for time 1, and slightly above for time 2. However, at both levels, the large standard deviations indicate that was also a wide variation in agreement amongst participants with the method creators.

Overall, there were many descriptor level codes where there was a lack of agreement between the participants and method creators; however, two consistent problems were identified. First, the contributing factors where the method creators had selected two codes to describe a factor were consistently associated with poor agreement, as participants tended to agree on one code but not the other. Second, agreement regarding contributing factors relating to activity equipment and resources was poor, as some participants assigned "lack of equipment/medication/documentation" to the actor responsible (e.g. activity participant or leader), rather than to the object.

Table 5 Descriptive statistics for validity scores: mean (M) Index of Concordance (%) between method creators and participants for system level and descriptor level codes for each incident at time 1 and time 2

	System level				Descriptor level			
	Time				Time			
	1		2		1		2	
Incident	M	SD	M	SD	M	SD	M	SD
1	79.8	10.0	79.4	9.2	65.3	10.3	70.9	12.9
2	73.7	8.4	79.9	9.0	68.4	9.8	71.5	8.8
3	73.1	12.1	70.2	10.5	63.5	18.2	65.2	14.2
4	69.1	10.5	71.7	7.7	73.3	9.1	75.8	11.5
Overall	73.9	10.3	75.3	9.1	67.6	11.9	70.8	11.9

4. Discussion

The aim of this study was to evaluate the reliability and validity of a classification scheme that was developed to support the use of Accimap by LOA practitioners. Reliability and validity was evaluated for practitioner coding at both the system level and descriptor level of the classification scheme. The results show that the classification scheme shows acceptable levels of intra-rater reliability at the system and descriptor levels when used by the same practitioner to code incident reports on two separate occasions. However, the classification scheme did not consistently achieve acceptable levels of inter-rater reliability at the system level when used by multiple practitioners to code the same incident reports, and inter-rater reliability at the descriptor level was poor. In terms of validity, on average there was an acceptable level of agreement between the system level codes selected by the method creators and practitioners; however, the classification scheme did not consistently achieve acceptable levels of validity at the descriptor level. In addition, the extent of agreement varied widely amongst practitioners. Overall, the results indicate that the classification scheme does not currently satisfy reliability and validity requirements, and that further work is required. The implications of the findings for the design and development of contributing factors classification schemes are discussed in the following sections.

In line with many other evaluations of contributing factor classification schemes (Baysari et al., 2011; Gordon et al., 2005; O'Connor, 2008; O'Connor, Walliser, & Philips, 2010; Olsen, 2011; Olsen & Shorrock, 2010), the results highlight the difficulty of designing a scheme that is reliable and useful (i.e. sufficiently detailed for coding specific contributing factors). In general, classification schemes with specific categories tend to achieve poor levels of reliability (Gordon et al., 2005; O'Connor, 2008; O'Connor et al., 2010; Olsen & Shorrock, 2010). From these studies, researchers have typically concluded that the number of categories should be reduced or that practitioners require “more training” in order to be able to use the scheme. Both of these options reduce the potentially usefulness of applying the method in practice, especially if the method creators need to provide end users with extensive face-to face training to achieve acceptable levels of reliability and validity (Stanton, 2016). This is often not practical, especially when linked to a research project and the funding expires.

Whilst it is important to examine how the levels of reliability and validity improves with more exposure to the method, the findings from this study suggest an initial compromise based on the intended context of application. Specifically, the results show that the most detailed level of the scheme has acceptable levels of reliability when a single analyst applies it on repeated occasions.

Therefore, it would be appropriate for a safety manager to use this level to analyse an organisation's incidents over time, and aggregate the findings to identify trends. However, if multiple analysts were involved in analysing incidents, then to maintain reliability, only the highest level of the classification scheme should be applied. In order to identify recurring problems, a single analyst could then apply the more detailed level of the scheme to code the incidents prior to aggregation and analysis. Similarly, the findings suggest that it would be appropriate to implement the highest level of the classification scheme within an incident reporting system. Again, additional coding at the more detailed level could then be undertaken by a single analyst prior to aggregation and analysis. In all the cases described above, aggregate analyses produced by a single analyst should be verified (i.e. checked) by other analysts to ensure that the conclusions drawn from the reports about the specific factors involved are appropriate, and disagreements resolved through discussion. This approach is consistent with the procedure often reported in published Accimap analyses to establish the validity of the analyses (i.e. Salmon et al., 2012; Underwood & Waterson, 2014).

Applying this approach to other classification schemes is, of course, dependent upon the availability of information on their reliability, validity and training requirements. The lack of evidence in this area, compared to the vast number of classification schemes that have been developed for incident coding, has been repeatedly noted (Beaubien & Baker, 2002; Mitchell, Williamson, Molesworth, & Chung, 2014; Olsen, 2011; Olsen & Shorrock, 2010), and is consistent with research on the reliability and validity of human factors methods more generally (Stanton, 2016; Stanton & Stevenage, 1998; Stanton & Young, 1999). The iterative evaluation process that we have used to develop the contributing factor classification scheme reported on in this paper represents an attempt to address this gap in the literature. The findings of this study, compared with previous evaluations of the same scheme (Goode et al., 2014b; Taylor et al., 2015), illustrates that reliability and validity can change dramatically over the course of development. It is also likely that reliability and validity will change in a positive direction based on end-user familiarity and experience levels. Despite this, few studies have examined reliability and validity levels over a significant period. Studies with significant periods between coding exercises (e.g. 12 – 24 months) are therefore recommended.

The findings from this study are also relevant to the on-going debate within the ergonomics literature regarding the relationship between validity, reliability and utility, and their relative importance in the development of ergonomics methods. On the one side, Stanton (Stanton, 2016; Stanton & Young, 1999, 2003) has repeatedly called for further studies evaluating the validity and reliability of ergonomics methods, as there is little point employing methods that do not satisfy this basic requirement (i.e. they are not useful because they are inaccurate). On the other side, others have argued that it may not always be possible or necessary to develop methods that meets all three requirements (e.g. Waterson, Clegg, & Robinson, 2014; Waterson et al., 2016). For example, Waterson et al. (2016) argues that the primary feature of Accimap is that it is relatively easy to use and provides an understanding of the factors which caused the accident (i.e. it is useful, but not necessarily accurate). The latter point of view is potentially reasonable if Accimap is only applied to analyse single incidents. However, the findings from this study illustrate that establishing satisfactory levels of reliability and validity is extremely important if the intention is to aggregate the findings across multiple incidents.

One caveat to the conclusions that can be drawn from this study is that the focus was on coding pre-existing incident reports. As noted in the introduction, one purpose of the scheme is to guide the collection of data during investigations, and subsequently support the understanding of the contributing factors involved. Further research is required to determine whether the scheme supports practitioners' understanding of incidents from a systems perspective during the

investigation process. The reliability and validity of the scheme would be completely different under these conditions, and this requires further investigation.

The limitations of the current study should also be acknowledged, which also suggest directions for further research. The key limitation is that the study was conducted by the method creators. While the study was designed so that there was no contact between the participants and the research team during the study period, or through feedback on the results, different results may be obtained by an independent research team. For example, a different research team may have made different decisions regarding the selection of incident reports. Second, reliabilty and validity was only evaluated for a small number of incident reports, which only covered a portion of the classification scheme. This limits the conclusions that can be drawn regarding the overall reliability and validity of the scheme. As noted in the method, the number of reports was limited so that the study would not place an unreasonable burden on participants. Future studies could overcome this problem by dividing the sample into sub-groups who all code different reports. Third, the scope of the study was constrained to specifically examine the selection of codes to describe contributing factors; intra-rater reliability is likely to be poorer under conditions where practitioners independently identify relevant contributing factors from incident reports. Overall, these limitations suggest the need to examine the reliability and validity of the classification scheme when it is used by practitioners to analyse and code their own incident reports, in the course of their normal work activities. This study was undertaken following the evaluation reported in this paper, and is reported in a separate manuscript (Goode, Salmon, Taylor, Lenné, & Finch, 2016).

In conclusion, this paper describes the research undertaken to develop and evaluate a domain-specific contributing factor classification scheme to support practitioners' application of Accimap to analyse LOA incidents. The results suggest the requirements for implementing the scheme within the LOA context with minimal training may need to be reconsidered. It seems unlikely that the scheme will ever show acceptable levels of inter-rater reliability and criterion-referenced validity under these conditions. The study clearly illustrates the challenges associated with designing a valid, reliable and useful classification scheme to support the application of Accimap within organisations.

5. Acknowledgements

This project was supported by funding from the Australia Research Council (ARC) in partnership with Australian Camps Association, Outdoor Council of Australia, The Outdoor Education Group, Sport and Recreation Victoria, Victorian YMCA Accommodation Services Pty Ltd, Outdoors Victoria, Outdoor Recreation Industry Council (Outdoors NSW), Outdoors WA, Outdoors South Australia, Queensland Outdoor Recreation Federation, Wilderness Escape Outdoor Adventures, Venture Corporate Recharge, and Christian Venues Association (LP110100037). Paul Salmon's contribution was funded through his current Australian Research Council Future Fellowship (FT140100681). Natassia Goode's contribution was funded through the University of the Sunshine Coast. Caroline Finch was supported by a NHMRC Principal Research Fellowship (ID: 565900). The Australian Centre for Research into Injury in Sport and its Prevention (ACRISP) is one of the International Research Centres for Prevention of Injury and Protection of Athlete Health supported by the International Olympic Committee (IOC).

6. References

Baber, C., & Stanton, N. A. (1994). Task analysis for error identification: a methodology for designing error-tolerant consumer products. *Ergonomics, 37*(11), 1923-1941. doi:https://doi.org/10.1080/00140139408964958

Baysari, M. T., Caponecchi, C., & McIntosha, A. S. (2011). A reliability and usability study of TRACEr-RAV: The technique for the retrospective analysis of cognitive errors e For rail, Australian

version. *Applied Ergonomics, 42*, 842 - 859. doi:https://doi.org/10.1016/j.apergo.2011.01.009

Baysari, M. T., McIntosh, M. S., & Wilson, J. R. (2008). Understanding the human factors contribution to railway accidents and incidents in Australia. *Accident Analysis & Prevention, 40*(5), 1750-1757. doi: https://doi.org/10.1016/j.aap.2008.06.013

Beaubien, J. M., & Baker, D. P. (2002). A review of selected aviation human factors taxonomies, accident/incident reporting systems and data collection tools. *International Journal of Applied Aviation Studies, 2*(2), 11-36.

Branford, K., Hopkins, A., & Naikar, N. (2009). Guidelines for AcciMap analysis *Learning from high reliability organisations*: CCH Australia Ltd.

Brookes, A. (2003). Outdoor education fatalities in Australia 1960-2002. Part 2. Contributing circumstances: supervision, first aid, and rescue. *Australian Journal of Outdoor Education, 7*(2), 34-42.

Brookes, A. (2004). Outdoor education fatalities in Australia 1960-2002. Part 3. Environmental circumstances. . *Australian Journal of Outdoor Education, 8*(1), 44-56.

Brookes, A., Smith, M., & Corkill, B. (2009). *Report to the Trustees of the Sir Edmund Hillary Outdoor Pursuit Centre of New Zealand: Mangatepopo Gorge Incident, 15th April 2008.* 5th September 2016 Retrieved from http://www.hillaryoutdoors.co.nz/newsite/wp-content/uploads/2013/06/091015-IRT-OPC_-Report.pdf

Chauvin, C., Lardjane, S., Morel, G., Clostermann, J.-P., & Langard, B. (2013). Human and organisational factors in maritime accidents: Analysis of collisions at sea using the HFACS. *Accident Analysis & Prevention, 59*(0), 26-37. doi:http://dx.doi.org/10.1016/j.aap.2013.05.006

Cornelissen, M., McClure, R., Salmon, P. M., & Stanton, N. A. (2014). Validating the Strategies Analysis Diagram: Assessing the reliability and validity of a formative method. *Applied Ergonomics, 45*(6), 1484-1494. doi:http://dx.doi.org/10.1016/j.apergo.2014.04.010

Curtis, R. (1995). *OA Guide to Outdoor Safety Management*. 5th September 2016 Retrieved from http://www.princeton.edu/~oa/files/safeman.pdf

Davenport, C. J. (2010). *Mangatepopo Coroners report*. 5th September 2016 Retrieved from Auckland: http://outdoorcouncil.asn.au/doc/Coroners_Report_OPC.pdf

Finch, C. F., Orchard, J. W., Twomey, D. M., Saleem, M. S., Ekegren, C. L., Lloyd, D. G., & Elliott, B. C. (2012). Coding OSICS sports injury diagnoses in epidemiological studies: does the background of the coder matter? *British Journal of Sports Medicine, 48*(7), 522-556. doi:https://doi.org/10.1136/bjsports-2012-091219

Fleishman, E. A., Quaintance, M. K., & Broedling, L. A. (1984). *Taxonomies of human performance: The description of human tasks*. U.S.A.: Academic Press.

Goode, N., Finch, C., Cassell, E., Lenne, M. G., & Salmon, P. M. (2014a). What would you like? Identifying the required characteristics of an industry-wide incident reporting and learning system for the led outdoor activity sector. *Australian Journal of Outdoor Education, 17*(2), 2–15.

Goode, N., Salmon, P. M., Lenne, M., & Finch, C. F. (2014b). A test of a systems theory-based incident coding taxonomy for risk managers. In P. Arezes & P. Carvalho (Eds.), *Advances in Safety Management and Human Factors* (Vol. 10 of Advances in Human Factors and Ergonomics 2014, pp. 5098-5108).

Goode, N., Salmon, P. M., Lenne, M., & Finch, C. F. (2014c). A test of a systems theory-based incident coding taxonomy for risk managers. *Advances in Safety Management and Human Factors, 10*, 5098-5108.

Goode, N., Salmon, P. M., Taylor, N. Z., Lenné, M. G., & Finch, C. F. (2016). Lost in translation: the validity of a systemic accident analysis method embedded in an incident reporting software tool. *Theoretical Issues in Ergonomics Science, 17*(5-6), 483-506. doi:https://doi.org/10.1080/1463922x.2016.1154230

Cite this paper: Goode, N., Salmon, P. M., Taylor, N. Z., Lenné, M. G., & Finch, C. F. (2017). Developing a contributing factor classification scheme for Rasmussen's AcciMap: Reliability and validity evaluation. *Applied Ergonomics*, 64, 14-26.

Gordon, R., Flin, R., & Mearns, K. (2005). Designing and evaluating a human factors investigation tool (HFIT) for accident analysis. *Safety Science, 43*(3), 147-171.

Haddock, C. (2004). *Outdoor safety : risk management for outdoor leaders*. Wellington N.Z.: New Zealand Mountain Safety Council

Hogan, R. (2002). The crux of risk management in outdoor programs--minimising the possibility of death and disabling injury. *Australian Journal of Outdoor Education, 6*(2), 71(76).

Kirwan, B. (1997). The validation of three human reliability quantification techniques — THERP, HEART and JHEDI: part iii — Practical aspects of the usage of the techniques. *Applied Ergonomics, 28*(1), 27-39. doi:http://dx.doi.org/10.1016/S0003-6870(96)00046-4

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*: Sage.

Leveson, N. (2004). A new accident model for engineering safer systems. *Safety Science, 42*(4), 237-270. doi:http://dx.doi.org/10.1016/S0925-7535(03)00047-X

Makeham, M. A., Stromer, S., Bridges-Webb, C., Mira, M., Saltman, D., Cooper, C., & Kidd, M. R. (2008). Patient safety events reported in general practice: a taxonomy. *Quality and Safety in Health Care, 17*(1), 53-57. doi:https://doi.org/10.1136/qshc.2007.022491

Mitchell, R. J., Williamson, A. M., Molesworth, B., & Chung, A. Z. Q. (2014). A review of the use of human factors classification frameworks that identify causal factors for adverse events in the hospital setting. *Ergonomics, 57*(10), 1443-1472. doi:10.1080/00140139.2014.933886

O'Connor, P. (2008). HFACS with an additional layer of granularity: validity and utility in accident analysis. *Aviation Space & Environmental Medicine, 79*(6), 599-606. doi:https://doi.org/10.3357/asem.2228.2008

O'Connor, P., Walliser, J., & Philips, E. (2010). Evaluation of a human factors analysis and classification system used by trained raters. *Aviat Space Environ Med, 81*(10), 957-960. doi:https://doi.org/10.3357/asem.2843.2010

Olsen, N. S. (2011). Coding ATC incident data using HFACS: Inter-coder consensus. *Safety Science, 49*(10), 1365-1370. doi:http://dx.doi.org/10.1016/j.ssci.2011.05.007

Olsen, N. S. (2013). Reliability studies of incident coding systems in high hazard industries: A narrative review of study methodology. *Applied Ergonomics, 44*(2), 175-184. doi:http://dx.doi.org/10.1016/j.apergo.2012.06.009

Olsen, N. S., & Shorrock, S. T. (2010). Evaluation of the HFACS-ADF safety classification system: Inter-coder consensus and intra-coder consistency. *Accident Analysis & Prevention, 42*(2), 437-444. doi:http://dx.doi.org/10.1016/j.aap.2009.09.005

Rasmussen, J. (1997). Risk management in a dynamic society: A modelling problem. *Safety Science, 27*(2/3), 183-213.

Ross, A. J., Wallace, B., & Davies, J. B. (2004). Technical note: measurement issues in taxonomic reliability. *Safety Science, 42*(8), 771-778. doi:http://dx.doi.org/10.1016/j.ssci.2003.10.004

Salmon, P. M., Cornelissen, M., & Trotter, M. J. (2012). Systems-based accident analysis methods: A comparison of Accimap, HFACS, and STAMP. *Safety Science, 50*(4), 1158-1170.

Salmon, P. M., Goode, N., Lenné, M. G., Finch, C. F., & Cassell, E. (2014). Injury causation in the great outdoors: A systems analysis of led outdoor activity injury incidents. *Accident Analysis & Prevention, 63*, 111-120. doi:http://dx.doi.org/10.1016/j.aap.2013.10.019

Salmon, P. M., Goode, N., Taylor, N. Z., Lenne, M. G., Dallat, C., & Finch, C. F. (2016). Rasmussen's legacy in the great outdoors: a new incident reporting and learning system for led outdoor activities. *Applied Ergonomics, 59*, 637-648. doi:10.1016/j.apergo.2015.07.017

Salmon, P. M., Williamson, A., Lenne, M., Mitsopoulos-Rubens, E., & Rudin-Brown, C. M. (2009). *The role of Human Factors in led outdoor activity incidents: Literature review and exploratory analysis*. 12th January 2017 Retrieved from Australia: http://outdoorcouncil.asn.au/wp-content/uploads/2016/08/OAI_REPORT_FINAL_VERSION_OCT_15th_2009.pdf

Salmon, P. M., Williamson, A., Lenne, M., Mitsopoulos-Rubens, E., & Rudin-Brown, C. M. (2010). Systems-based accident analysis in the led outdoor activity domain: Application and

evaluation of a risk management framework. *Ergonomics, 53*(8), 927-939. doi:10.1080/00140139.2010.489966

Service Skills Australia. (2010). *National Outdoor Sector Survey 2010: Quantifying the outdoor workforce*. 01.11.2016 Retrieved from Australia: http://www.qorf.org.au/wp-content/uploads/2014/03/NOSS2010_Report.pdf

Service Skills Australia. (2013). *2013 National Outdoor Sector Survey*. 01.11.2016 Retrieved from Australia: http://qorf.org.au/wp-content/uploads/2014/08/NOSS_2013_Report.pdf

Shappell, S. A., & Wiegmann, D. A. (2012). *A human error approach to aviation accident analysis: The human factors analysis and classification system*. U.K.: Ashgate Publishing, Ltd.

Stanton, N., & Young, M. (1998). Is utility in the mind of the beholder? A study of ergonomics methods. *Applied Ergonomics, 29*(1), 41-54. doi:https://doi.org/10.1016/s0003-6870(97)00024-0

Stanton, N. A. (2016). On the reliability and validity of, and training in, ergonomics methods: a challenge revisited. *Theoretical Issues in Ergonomics Science, 17*(4), 345-353. doi:10.1080/1463922X.2015.1117688

Stanton, N. A., Salmon, P., Harris, D., Marshall, A., Demagalski, J., Young, M. S., . . . Dekker, S. (2009). Predicting pilot error: Testing a new methodology and a multi-methods and analysts approach. *Applied Ergonomics, 40*(3), 464-471. doi:http://dx.doi.org/10.1016/j.apergo.2008.10.005

Stanton, N. A., & Stevenage, S. V. (1998). Learning to predict human error: issues of acceptability, reliability and validity. *Ergonomics, 41*(11), 1737-1756. doi:10.1080/001401398186162

Stanton, N. A., & Young, M. S. (1999). What price ergonomics? *Nature, 399*(6733), 197-198.

Stanton, N. A., & Young, M. S. (2003). Giving ergonomics away? The application of ergonomics methods by novices. *Appl Ergon, 34*(5), 479-490. doi:10.1016/s0003-6870(03)00067-x

Taylor, N. Z., Goode, N., Salmon, P. M., Lenne, M. G., & Finch, C. F. (2015). *Which code is it? Inter-rater reliability of systems theory-based causal factor taxonomy for the outdoor sector*. Paper presented at the 19th Triennial Congress of the International Ergonomics Association, Melbourne, Australia.

Underwood, P., & Waterson, P. (2013a). Accident Analysis Models and Methods: Guidance for Safety Professionals. Retrieved from https://dspace.lboro.ac.uk/dspace-jspui/bitstream/2134/13865/4/Underwood%20and%20Waterson%20(2013)%20-%20Accident%20Analysis%20Models%20and%20Methods%20-%20Guidance%20for%20Safety%20Professionals.pdf

Underwood, P., & Waterson, P. (2013b). Systemic accident analysis: Examining the gap between research and practice. *Accident Analysis & Prevention, 55*, 154-164. doi:http://dx.doi.org/10.1016/j.aap.2013.02.041

Underwood, P., & Waterson, P. (2014). Systems thinking, the Swiss Cheese Model and accident analysis: A comparative systemic analysis of the Grayrigg train derailment using the ATSB, AcciMap and STAMP models. *Accident analysis and prevention, 68*, 75-94.

Walker, P. B., O'Connor, P., Phillips, H. L., Hahn, R. G., & Dalitsch, W. W. (2011). Evaluating The Utility of DoD Hfacs Using Lifted Probabilities. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 55*(1), 1793-1797. doi:10.1177/1071181311551372

Wallace, B., & Ross, A. J. (2006a). Appendix: Carrying out a reliability trial *Beyond Human Error* (pp. 243-245). U.S.A.: CRC Press.

Wallace, B., & Ross, A. J. (2006b). *Beyond Human Error*. U.S.A.: CRC Press.

Waterson, P., Clegg, C. W., & Robinson, M. (2014). Trade-offs between reliability, validity and utility in the development of human factors methods. *Human Factors in Organizational Design and Management XI, edited by O. Broberg, N. Fallentin, P. Hasle, PL Jensen, A. Kabel, ME Larsen, and T. Weller. Santa Monica, CA: IEA Press.(CD ROM)*.

Cite this paper: Goode, N., Salmon, P. M., Taylor, N. Z., Lenné, M. G., & Finch, C. F. (2017). Developing a contributing factor classification scheme for Rasmussen's AcciMap: Reliability and validity evaluation. *Applied Ergonomics*, 64, 14-26.

Waterson, P., Jenkins, D. P., Salmon, P. M., & Underwood, P. (2016). 'Remixing Rasmussen': The evolution of Accimaps within systemic accident analysis. *Applied Ergonomics, 59*, 483-503. doi:https://doi.org/10.1016/j.apergo.2016.09.004

Wiegmann, D. A., & Shappell, S. A. (2001). Human error analysis of commercial aviation accidents: Application of the Human Factors Analysis and Classification System (HFACS). *Aviation, space & environmental medicine, 72*(11), 1006-1016. doi:https://doi.org/10.1037/e420582004-001

Williams, I., & Allen, N. (2012). *National Survey of Australian Outdoor Youth Programs*. 01.11.2016 Retrieved from http://www.oypra.org.au/resources/OYPRA_Australian_Outdoor_Survey_Report_2012.pdf